A Single-Server Priority Queueing-Location Model

Rajan Batta
Department of Industrial Engineering, State University of New York at Buffalo, Buffalo, New York 14260

Richard C. Larson and Amedeo R. Odoni
Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

This paper considers the problem of locating a single server on a network, relaxing the assumption that the server is always available for service, and explicitly accounting for queueing. The resulting queueing-location model allows for an arbitrary number of priority classes. Properties of the objective function are developed and algorithms presented for obtaining the optimal location on tree and cyclic networks. Sensitivity analysis with respect to the average arrival rate of calls is investigated. A numerical example is presented to illustrate the results of this paper. The major conclusions of the paper include: (a) the optimal location need not be at a node of the network, (b) the optimal location changes as a function of the arrival rate of calls into the system, (c) the optimal location is usually different from that obtained by grouping all calls into one priority class.

INTRODUCTION

In recent papers Berman, Larson, and Chiu [3] and Chiu, Berman, and Larson [5] have proposed and analyzed a queueing-location model, which they refer to as the Stochastic Queue Median (SQM) model. SQM assumes that if calls are waiting in queue to be served, the dispatcher, when picking the next call to be serviced, will consider all calls to be of the same importance and will choose the next call by a rule that depends only upon the relative position of the calls in queue. For example, SQM allows for queueing disciplines like First-Come-First-Served (FCFS), Last-Come-First-Served (LCFS) and Service-In-Random-Order (SIRO), etc.

This type of priority rule is clearly inappropriate in many contexts. For example, in urban emergency services, calls that involve danger to human life (e.g., a major accident in the case of an ambulance system or a "violent crime in progress" in the

NETWORKS, Vol. 8 (1988) 87–103

CCC 0028–3045/88/020087-17$04.00

case of police patrols) deserve and receive higher priority over calls for more routine incidents. A realistic model for such contexts should allow for prioritization of calls for service.

Our aim in this paper is to formulate and provide solution techniques for a single server queueing-location model which allow calls to be selected from an arbitrary number of priority classes. The exact number of priority classes which should be employed depends on the particular application at hand and requires study in its own right.

The reason for restricting our attention to a single server model is that, in the course of our analysis, we shall need a closed-form expression from queueing theory for the average waiting time in queue for calls of various priorities. While such an expression does exist for an M/G/1 queueing system, as of today, none exists for an M/G/m system when m (the number of servers) is greater than one. Two attempts to solve multiple server queueing-location models through an approximation technique are the papers by Berman and Larson [2] and Larson [9].

We refer to the model developed in this paper as a K-Priority Queueing-Location (K-PQL) model. The principal results presented in this paper can be summarized as follows:

1. The optimal K-PQL model location is usually different from that obtained by grouping calls from all priorities into a single category and using the SQM model developed in [3].
2. For a tree network we establish properties of the objective function that enable one to trim off portions of the tree where the optimal location cannot lie.
3. We characterize the set of optimal locations for extreme values of the arrival rate. For a tree network these exist at unique points. For a cyclic network the set has finite cardinality. Elements of the set can be efficiently identified in both cases.

FORMULATION

Given an undirected network $G = (N,A)$, with N being the node set, $|N| = n$, A being the arc set, $|A| = a$, assume that travel between two points x and y on G occurs on a shortest length path connecting these two points and denote the length of the path by $d(x,y)$.

Calls arrive solely at the nodes of the network. They are divided into K priority classes, indexed by k, $k = 1,2, \ldots, K$; the lower the value of k, the higher the priority of the call. Arrivals of calls of priority k constitute arrivals from a time homogeneous Poisson stream, with average arrival rate λ_k. Given that an incoming call is of priority k, the *a priori* probability that the call originated from node i is f_{ik}. We also define $\lambda = \sum_{k=1}^{K} \lambda_k$ and $F_k = \lambda_k/\lambda$.

The service system operates as follows. The server is located at a point on the network, called the "home location," when not busy servicing a call. After servicing a call the server always returns to the home location, before answering the next call. If a call enters the system, and the server is busy serving another call, the arriving call is entered into a queue. The queue is depleted by employing a Head-

of-the-Line (HOL) non-preemptive priority queueing discipling, FCFS within each priority class.

The system time of a call can be divided into two portions: (a) queueing delay; and (b) service time (travel time to call, on-scene service time at scene of call, travel time from call, and off-scene service time at home location).

The server, when travelling to a priority k call, travels with constant speed v_k. Furthermore, when he travels from the scene of the priority k call to his home location, he travels with speed $v_k/(\beta - 1)$, $\beta > 1$, where β is a prespecified constant.

The first and second moments of *nontravel time related service time* (on-scene plus off-scene) for a priority k call from node i, are denoted by finite constants $\overline{W}_{ik}$ and $\overline{W^2_{ik}}$, respectively. In particular, we note that these are assumed independent of the home location of the server.

To optimize for the "best" home location of the server we have to decide upon an objective function. In an urban emergency service system and for a priority k call, quality of service if often measured with reference to *Response Time, TR_k*. When the home location is at x on the network,

$$TR_k(x) = Q_k(x) + t_k(x), \tag{1}$$

where

$Q_k(x)$ = queueing delay of a priority k call, when the server's home location is at x, and

$t_k(x)$ = travel time to the priority k call, when the server's home location is at x.

$TR_k(x)$, $Q_k(x)$, and $t_k(x)$ are random variables, as a result of the spatial and temporal uncertainty associated with the calls for service. Denoting the average of a random variable by y by $\bar{y}$, and attaching an *importance factor* $\pi_k(\pi_k \geqslant 0)$ to priority k calls, the *weighted average response time* to a call, when the server's home location is at x, is given by:

$$\overline{TR}(x) = \sum_{k=1}^{K} \pi_k \overline{TR}_k(x). \tag{2}$$

We wish to find a point x which minimizes $\overline{TR}(x)$.

ANALYSIS

To simplify presentation, the following notation shall be used. A variable $A_k(x)$ indicates the variable A depends on (i) the priority class k, and (ii) the home location x of the server. A variable $B(x)$ indicates that variable B depends on the home location x of the server.

Denoting the expected travel time to a call, expected service time and second moment of service time by the symbols $\bar{t}$, $\overline{S}$, and $\overline{S^2}$, respectively, we have in terms of variables previously introduced:

$$\bar{t}_k(x) = \sum_{i \in N} f_{ik}\, d(x,i)/v_k, \tag{3}$$

$$\bar{S}_k(x) = \sum_{i \in N} f_{ik}\, [\overline{W}_{ik} + \beta d(x,i)/v_k], \tag{4}$$

$$\overline{S_k^2}(x) = \sum_{i \in N} f_{ik}[\overline{W_{ik}^2} + 2\overline{W}_{ik}\beta d(x,i)/v_k + (\beta d(x,i)/v_k)^2], \tag{5}$$

$$\overline{S^2}(x) = \sum_{k=1}^{K} F_k \overline{S_k^2}(x). \tag{6}$$

The assumption that the server always returns to his home location before answering successive calls leads to independent and identically distributed (i.i.d.) service times. The system operates as an M/G/1 queueing system with K priority classes, nonpreemptive priorities, FCFS within each priority class.

The average queueing delay for priority k calls, denoted by symbol $\overline{Q}$, is (for example, see Kleinrock [8]):

$$\overline{Q}_k(x) = \begin{cases} \dfrac{\frac{\lambda}{2}\overline{S^2}(x)}{\left[1 - \sum_{j=1}^{k} \lambda_j \bar{S}_j(x)\right]\left[1 - \sum_{j=1}^{k-1} \lambda_j \bar{S}_j(x)\right]}, & \text{if } 1 - \sum_{j=1}^{k} \lambda_j \bar{S}_j(x) > 0, \\ +\infty, & \text{otherwise.} \end{cases} \tag{7}$$

Using (1)–(7) we can therefore write an expression for the weighted average response time, $\overline{TR}(x)$.

In order to search efficiently for the minimum of $\overline{TR}(x)$, we identify regions of the network where $\overline{TR}(x)$ takes on a unique functional form. We briefly outline here the approach used, as it is virtually identical to the one in [3] and Chiu [4]. The weighted travel distance function is piecewise linear and concave over any arc of a network. We will define the region of an arc over which this weighted distance function is linear as a primary region. Further, various associated functions (when finite) are differentiable over the interior of a primary region.

The steps in identifying a minimum of $\overline{TR}(x)$ on the network are:

(1) Identify the primary regions on the network.

(2) Eliminate primary regions (for a tree network) where the minimum of $\overline{TR}(x)$ cannot lie.

(3) Find a local minimum for each of the remaining primary regions. A descent algorithm (for example see Shapiro [10]) can be used for this purpose, after the feasible

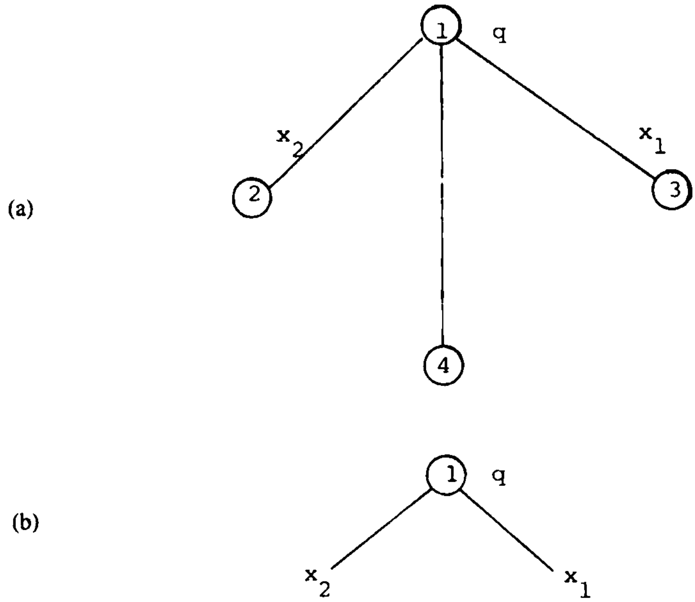

FIG. 1. Illustration for proof of Theorem 3: (a) Graph G, (b) Subgraph G'.

portion of the primary region has been identified. The feasible portion of a primary region is the section on which $\overline{TR}(x)$ is finite.

(4) Compare these local minima to obtain the home location x on G.

SIMPLIFICATION FOR TREES

In contrast to the case of cyclic networks, we can establish properties of the objective function on trees which allow us to eliminate primary regions or portions thereof, where the minimum of $\overline{TR}(x)$ cannot lie. This is the purpose of this section.

Theorem 1. $\overline{Q}_k(x)$ is strongly quasiconvex (when finite) on any path of a tree network, for $\lambda > 0$, $\forall\, k = 1,2,\ldots,K$.

Proof. See Batta [1]. ■

Theorem 2. $\sum_{k=1}^{K} \pi_k \bar{t}_k(x)$ is convex on any path of a tree.

Proof. Convexity of $\bar{t}_k(x)$ on any path of a tree is well known, see Goldman [6]. The result follows from the fact that $\pi_k \geqslant 0$, $\forall\, k = 1,2,\ldots,K$. ■

We use the results stated in Theorems 1 and 2 to eliminate portions of a tree network where the global minimum for $\overline{TR}(x)$ cannot lie.

Theorem 3. Let $x_1, x_2, \ldots, x_K$ be the unique points on the tree network G which respectively minimize $\overline{Q}_1(x), \overline{Q}_2(x), \ldots, \overline{Q}_K(x)$. Let q be any point which minimizes $\sum_{k=1}^{K} \pi_k \bar{t}_k(x)$. Finally, let G' be the smallest connected subgraph of G which contains all the points, $x_1, x_2, \ldots, x_K$ and q. Then the minimizer of $\overline{TR}(x)$ lies on subgraph G'.

Proof. The proof is by contradiction. Suppose this were not the case. Then for a tree network G (Fig. 1a) and associated subgraph G' (Fig. 1b) a minimizer of $\overline{TR}(x)$ lies at a point $x^* \notin G'$, as shown in Figure 1a.

By moving the location from x^* to the nearest point on subgraph G' (node q in Fig. 1a), $\overline{Q}_1(x), \overline{Q}_2(x), \ldots, \overline{Q}_K(x)$ all strictly decrease, by Theorem 1. Also $\sum_{k=1}^{K} \pi_k \bar{t}_k(x)$ cannot increase, by Theorem 2. From (1), (2) and from the fact that $\pi_k \geq 0\ \forall\ k = 1,2, \ldots, K$, it follows that $\overline{TR}(x)$ must strictly decrease through this movement. The Theorem follows. ∎

To complete our discussion on elimination of primary regions of a tree network G where the minimum of $\overline{TR}(x)$ cannot lie, we state algorithms for finding:

(a) A minimizer of $\sum_{k=1}^{K} \pi_k \bar{t}_k(x)$.

(b) Minimizers of $\overline{Q}_1(x), \overline{Q}_2(x), \ldots, \overline{Q}_K(x)$.

The minimizers found in (a) and (b) can then be used to form the subgraph G' of G to which we can restrict our search.

Algorithm for Finding Minimizer of $\sum_{k=1}^{K} \pi_k \bar{t}_k(x)$

A nodal minimizer of $\sum_{k=1}^{K} \pi_k \bar{t}_k(x)$ can be found by using Goldman's algorithm for finding the one median on a tree, see Goldman [6], with the weight of node i being

$$w_i = \sum_{k=1}^{K} \pi_k f_{ik}/v_k.$$

Algorithm for Finding Minimizer of $\overline{Q}_k(x)$, $k = 1,2, \ldots, K$

By Theorem 1, $\overline{Q}_k(x)$ is strongly quasiconvex (when finite) on any path of a tree, $\forall\ k = 1,2, \ldots, K$. An algorithm very similar to that used in [5] (for finding the SQM on a tree) can be developed. We omit the details.

SENSITIVITY ANALYSIS WITH RESPECT TO ARRIVAL RATE

In this section, we address the question: How does the optimal location of the facility change as the arrival rate of calls is increased from 0 to the maximum possible arrival rate, λ_{max} (λ_{max} is the *minimum* value of λ which renders *every* point of the network infeasible as a home location; specifically, if node j is a minimizer of $\sum_{k=1}^{K} \lambda_k \overline{S}_k(x)$, then $\lambda_{max} = 1/(\sum_{k=1}^{K} F_k\, \overline{S}_k(j))$. We formalize our observations in Theorem 4.

Theorem 4. Define sets S, T, S_1 and T_1 as follows:

$$S = \left\{x \mid x \in G, \sum_{k=1}^{K} \pi_k \bar{t}_k(x) \leq \sum_{k=1}^{K} \pi_k \bar{t}_k(y), \forall\, y \in G\right\},$$

$$T = \left\{x \mid x \in S, \overline{S^2}(x) \leq \overline{S^2}(y), \forall\, y \in S\right\},$$

$$S_1 = \left\{x \mid x \in G, \sum_{k=1}^{K} \lambda_k \bar{S}_k(x) \leq \sum_{k=1}^{K} \lambda_k \bar{S}_k(y), \forall\, y \in G\right\},$$

$$T_1 = \left\{x \mid x \in S_1, \frac{\overline{S^2}(x)}{\left[1 - \lambda_{\max} \sum_{k=1}^{K-1} F_k\, \bar{S}_k(x)\right]} \leq \frac{\overline{S^2}(y)}{\left[1 - \lambda_{\max} \sum_{k=1}^{K-1} F_k\, \bar{S}_k(y)\right]}, \forall\, y \in S_1\right\}.$$

Then, the following hold:

(a) for every $x_0 \notin T$ and $y_0 \in T$, one can find $\varepsilon_1 > 0$ such that $\overline{TR}(x_0) > \overline{TR}(y_0)$ whenever $0 \leq \lambda \leq \varepsilon_1$,

(b) for every $x_0 \notin T_1$ and $y_0 \in T_1$, one can find $\varepsilon_2 > 0$ such that $\overline{TR}(x_0) > \overline{TR}(y_0)$ whenever $\lambda_{\max} - \varepsilon_2 \leq \lambda < \lambda_{\max}$, and

(c) the sets T and T_1 have finite cardinality. When G is a tree T and T_1 are singleton sets.

Proof. See Appendix. ∎

To present intuitively the contents of Theorem 4 we distinguish two kinds of Hakimi medians. A point is a type I Hakimi median, denoted by HM1, if it minimizes $\sum_{k=1}^{K} \pi_k \bar{t}_k\,(x)$ on G. Similarly, a point is a type II Hakimi median, denoted by HM2, if it minimizes $\sum_{k=1}^{K} \lambda_k \bar{S}_k\,(x)$ on G. HM1 is the set of points which minimizes the weighted (by priority class) average travel time to a call, and HM2 is the set of points which minimizes the system average service time (or system average travel time) to a call.

If HM1 *is* unique, then from [7] is exists at a node j of the network (the fact that π_k's are not nodal weights does not change this observation). In this case, Theorem 4 asserts that node j is the optimal location for λ close to 0. If HM1 is *not* unique, then the optimal location is a point in the set of points HM1, which minimizes $\overline{S^2}(x)$.

If HM2 *is* unique, then from [7] is exists at a node k of the network. In this case, Theorem 4 asserts that node k is the optimal location for λ close to $\lambda_{\max}$. If HM2 is *not* unique, then the optimal location is a point in the set of points HM2, which minimizes a function of $\overline{S^2}(x)/D$, where D depends only on the calls of the first $K - 1$ priorities.

Theorem 4 also asserts that the optimal location is *unique* when λ is close to either

0 or $\lambda_{\max}$ for the case of a tree network, and the set of optimal locations is finite for a cyclic network.

To find the elements of T and T_1 on a tree network we can exploit the strict convexity of $\overline{S^2}(x)$ and

$$\frac{\overline{S^2}(x)}{\left[1 - \lambda_{\max}\sum_{k=1}^{K-1} F_k \, \bar{S}_k(x)\right]}$$

on *any* path of a tree, to develop algorithms similar to that of finding the minimizer of $\bar{Q}_k(x)$ on a tree.

To find elements of T and T_1 respectively on a cyclic network we can exploit the strict convexity of the above mentioned functions on a primary region.

For intermediate values of λ, the average queueing delay of a call and the average travel time to a call are of comparable magnitude. The second moment of service time plays a central role in determination of the optimal location. Our computational experience with the model indicates similar behavior to that observed in [3] and [5].

NUMERICAL EXAMPLE

In this section we illustrate the results of this paper through a simple numerical example.

We use the problem instance described in Figure 2. The first column of Table I displays the optimal location when λ varies from 0.045 to 0.855 under the 2-priority queueing-location (2-PQL) model. The second column of Table I displays the corresponding optimal location obtained by merging the calls from priorities 1 and 2 into one category, and using the SQM model in [3]. The third column of Table I gives the corresponding optimal location when ignoring queueing and using the model in [7], which we refer to as the Hakimi model.

Table II displays the objective function value obtained *under the 2-PQL model* when the facility is located at (i) the optimal 2-PQL model location, (ii) the optimal SQM model location, and (iii) the optimal Hakimi model location. Define:

$$A(x) = \pi_1 \bar{t}_1(x) + \pi_2 \bar{t}_2(x),$$

$$B(x) = F_1 \bar{S}_1(x) + F_2 \bar{S}_2(x).$$

A. We note that $A(x)$, the weighted sum of average travel times to calls of priorities 1 and 2, is uniquely minimized at node 1.

B. $B(x)$, the system average service time to a call (of priority 1 or 2), is minimized at *any* point on arc (3,5).

C. $\overline{S^2}(x)$, the second moment of the service time of all calls (of priority 1 or 2) is minimized uniquely at a point 2.5 units from node 3, on arc (3,5). Here we have an unusual situation: the point which minimizes the second moment of service time also minimizes average service time.

D. $\lambda_{\max} = 1/(F_1 \bar{S}_1(x) + F_2 \bar{S}_2(x)) \Big|_{x \text{ is on arc } (3,5)}$

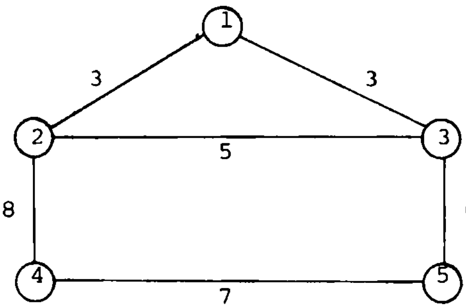

FIG. 2. Numerical example.

Number(s) next to arc(s) represent arc lengths. $\beta = 2$; travel to a call is with same speed as travel from a call. $\pi_1 = 2$; $\pi_2 = 1$; importance factors for priority 1 and 2 calls respectively. $P_1 = [1/2, 1/8, 1/8, 1/8, 1/8]$; i th entry is probability that the next priority 1 call comes from node i. $P_2 = [1/8, 1/8, 1/8, 1/8, 1/2]$; i th entry is probability that the next priority 2 call comes from node i. $v_1 = v_2 = 10$; the travel speeds for priority 1 and 2 calls respectively. $W_1 = [0.1, 0.1, 0.1, 0.1, 0.1]$; i th entry is mean non-travel time related service time for priority 1 calls from node i. $W_2 = [0.1, 0.1, 0.1, 0.1, 0.1]$; i th entry is mean non-travel time related service time for priority 2 calls from node i. $W_1^2 = [0.01, 0.01, 0.01, 0.01, 0.01]$; i th entry is second moment of non-travel time related service time for priority 1 calls from node i. $W_2^2 = [0.01, 0.01, 0.01, 0.01, 0.01]$; i th entry is second moment of non-travel time related service time for priority 2 calls from node i. $F_1 = 1/3$, $F_2 = 2/3$; fraction of calls of priority 1 and 2 respectively.

Further define:

$$C(x) = \overline{S^2}(x)/(1 - \lambda_{\max} F_1 \bar{S}_1(x)).$$

Then $C(x)$, when x is on arc (3,5), is minimized uniquely at a point 1.77 units from node 3, on arc (3,5). From the assertion in Theorem 4(b), the optimal 2-PQL location is at the point which minimizes $C(x)$ on arc (3,5), when λ is sufficiently close to $\lambda_{\max}$.

Now we are in a position to discuss the results in Tables I and II.

In our example the minimizer of $\overline{S^2}(x)$ also minimizes $B(x)$, therefore we are assured

TABLE I. Optimal location under (a) Hakimi Model, (b) SQM Model, and (c) 2-PQL Model.

λ	2-PQL Model	SQM Model (priorities grouped together)	Hakimi Model
0.045	node 1	arc (3,5), 2.5 from node 3	any point on arc (3,5)
0.180	node 1	arc (3,5), 2.5 from node 3	any point on arc (3,5)
0.360	arc (3,5), 0.995 from node 3	arc (3,5), 2.5 from node 3	any point on arc (3,5)
0.540	arc (3,5), 1.451 from node 3	arc (3,5), 2.5 from node 3	any point on arc (3,5)
0.720	arc (3,5), 1.598 from node 3	arc (3,5), 2.5 from node 3	any point on arc (3,5)

TABLE II. Objective function value when using 2-PQL Model assumptions and locating facility as in Table I.

λ	2-PQL Model	SQM Model (priorities grouped together)	Hakimi Model (assuming location is at node 3)
0.045	1.461	1.760	1.589
0.180	1.973	2.148	2.031
0.360	2.813	2.882	2.842
0.540	4.117	4.186	4.242
0.720	7.827	7.955	8.199

that for *all* feasible values of λ ($0 < \lambda < \lambda_{max}$), the optimal location under SQM will be a distance 2.5 from node 3, on arc (3,5).

For the deterministic case the optimal location is *any* point on arc (3,5); we have chosen node 3 arbitrarily.

For low values of λ, the 2-PQL model favors node 1 as an optimal location primarily because of the concern reflected for priority 1 calls ($\pi_1 = 2$, $\pi_2 = 1$).

For intermediate values of λ, queueing delays and travel times to calls are of comparable magnitude: hence the optimal location for the 2-PQL model shifts towards the minimizer of the second moment of service time. We note that the higher importance attached to priority 1 calls tends to keep the location toward node 1, hence the optimal location is closer to node 1 than under the SQM model.

For large values of λ, we appeal to Theorem 2(b) to yield the optimal location of the facility under the 2-PQL model.

CONCLUDING REMARK

We end this paper by commenting on properties of the algorithms developed therein. $\overline{TR}(x)$ is finite if $\overline{Q}_1(x), \overline{Q}_2(x), \ldots, \overline{Q}_K(x)$ are finite. From Lemma 1 (in Appendix) we know that the $\overline{Q}_k(x)$'s are strongly quasiconvex (when finite) on a primary region of the network. Even though $\overline{TR}(x)$ is a positively weighted sum of the $\overline{Q}_k(x)$'s this does *not* necessarily imply any convexity property of $\overline{TR}(x)$ (when finite).

The strong quasiconvexity (when finite) of $\overline{TR}(x)$ persists in numerical examples because (see proof of Lemma 1 in Appendix) the *only* case when $\overline{Q}_k(x)$ may not be strictly convex (when finite) in a primary region is when its denominator is strictly convex; this happens when the mean service times with respect to calls of the first $k - 1$ and k priority classes are uniquely minimized at the same end of the primary region.

The authors would like to thank the anonymous referees for their helpful comments on an earlier version of this paper. This research was supported, in part, by the National Science Foundation, viz-a-vis Grant No. ECS-8204318; the support is gratefully acknowledged.

APPENDIX

This Appendix contains proofs of some of the theorems and lemmas in this paper. In order to prove Theorem 1 we need to prove a Lemma on strong quasiconvexity of $\overline{Q}_k(x)$ on a primary region of a network.

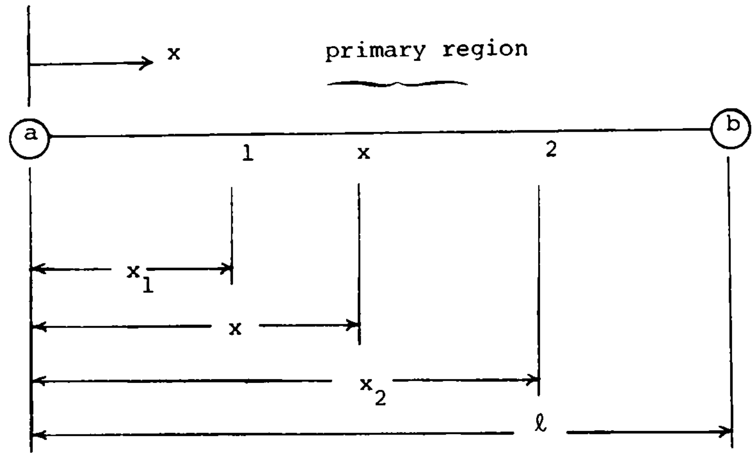

FIG. 3. Illustration used in proof of Lemma 1.

Lemma 1. $\overline{Q}_k(x)$ is strongly quasiconvex (when finite) on a primary region of a (cyclic or tree) network, for $\lambda > 0$, $\forall\, k = 1,2, \ldots, K$.

Proof. Consider a primary region 1–2 on arc (a,b) of the network, as shown in Figure 3.

We wish to prove strong quasiconvexity of $\overline{Q}_k(x,)$, $\lambda > 0$, for $k = 1,2, \ldots, K$, when the function is finite.

For notational convenience we shall suppress the parameter x in the formulae of $\overline{S}_k(x)$, $\overline{S_k^2}(x)$, $\bar{t}_k(x)$ and $\overline{S^2}(x)$, labeling them as $\overline{S}_k$, $\overline{S_k^2}$, $\bar{t}_k$, and $\overline{S^2}$, respectively, whenever the context leaves no ambiguity. In addition, we shall denote first and second derivatives of a function $f(x)$ by $f'(x)$ and $f''(x)$; further suppressing x, we write these as f' and f'', respectively.

We divide the analysis into two cases:

Case I; $k = 1$. In this case, the denominator of $\overline{Q}_1(x)$ is a linear function of x. It follows from [3] that $\overline{Q}_1(x)$ is strictly convex in a primary region.

Case II: $k > 1$.

$$\overline{Q}_k(x) = \frac{(\lambda/2)\overline{S^2}}{\left[1 - \sum_{j=1}^{k-1} \lambda_j \overline{S}_j\right]\left[1 - \sum_{j=1}^{k} \lambda_j \overline{S}_j\right]}. \tag{8}$$

Let $N(x) = (\lambda/2)\,\overline{S^2} = a + bx + cx^2$, and

$$D(x) = \left[1 - \sum_{j=1}^{k-1} \lambda_j \overline{S}_j\right]\left[1 - \sum_{j=1}^{k} \lambda_j \overline{S}_j\right] = (a_1 + b_1x)(a_2 + b_2x)$$
$$= d + ex + fx^2,$$

where $a,b,c,d,e,$ and f are constants that can be evaluated for a given primary region.

Since $\lambda > 0$, it follows from [3] that: $N(x) > 0 \ \forall\ x \in R \Rightarrow b^2 - 4ac < 0 \Rightarrow (N')^2 - 2N''N < 0$. We note that N, N', and N'' are defined for $x = 0$.

Also, $D(x)$ has real roots, because $D(x) = (a_1 + b_1x)(a_2 + b_2x)$.

Direct differentiation of $\overline{Q}_k(x)$ yields: $\overline{Q}_k''(x) = \{D\}^{-3}\{N''D^2 - 2D'DN' + 2(D')^2N - ND''D\}$.

Since $\overline{Q}_k(x)$ is assumed finite, it takes on the sign of:

$$T(x) = N''D^2 - 2D'DN' + 2(D')^2N - ND''D.$$

We distinguish 3 cases.

Subcase (a): $D'' = 0$. In this case $D(x)$ is linear in x and the arguments of Case I apply.

Subcase (b): $D'' < 0$. In this case, we can write:

$$\begin{aligned} T(x) &= N''D^2 - 2D'DN' + 2(D')^2N - ND''D \\ &\geqslant N''D^2 - 2D'DN' + 2(D')^2N, \end{aligned}$$

where the inequality follows from $N,D > 0$ and $D'' < 0$. The proof that $T(x) \geqslant 0 \ \forall\ x \in R$ follows directly from Case I, because D'' has been effectively removed from the function, and the proof is now identical to that presented in [3].

Subcase (c): $D'' > 0$. We distinguish two cases, when the roots of $D(x)$ are distinct, and when they are not distinct. Throughout this discussion we shall temporarily assume that $\overline{Q}_k(x)$ is defined by (8) even if $D(x) \leqslant 0$.

Sub-subcase (i): roots of $D(x)$ are distinct (Fig. 4). Let the roots of $D(x)$ be y_1 and y_2, respectively. In the interval (y_1,y_2), $D(x)$ is negative. Therefore, $\overline{Q}_k''(x)$ takes on the sign of $-T(x)$.

$$\begin{aligned} -T(x) &= -N''D^2 + 2D'DN' - 2(D')^2N + ND''D \\ &\leqslant -N''D^2 + 2D'DN' - 2(D')^2N, \end{aligned}$$

where, the last inequality follows from $N,D'' > 0$ and $D < 0$.

$$\begin{aligned} \text{Let } G(x) &= -N''D^2 + 2D'DN' - 2(D')^2N \\ &= -N''D^2 - [2D'DN' + 2(D')^2N]. \end{aligned}$$

It follows that $G(x) < 0 \ \forall\ x \in R$, from the arguments in subcase (b). Therefore $\overline{Q}_k(x)$ is *strictly concave* in x, when $x \in (y_1,y_2)$.

$\overline{Q}_k'(x) = 0$, yields a quadratic in x (the cubic terms cancel out). Since $\overline{Q}_k(x) \rightarrow -\infty$ when $x \rightarrow y_1$ from above, and $\overline{Q}_k(x) \rightarrow -\infty$ when $x \rightarrow y_2$ from below, strict concavity of $\overline{Q}_k(x)$ in the interval (y_1,y_2) implies that *exactly one* of the solutions to $\overline{Q}_k'(x) = 0$ lies in the interval (y_1,y_2). Therefore, exactly one solution to $\overline{Q}_k'(x) = 0$ lies in $R - (y_1,y_2)$, where $R = (-\infty, +\infty)$. This establishes strong quasiconvexity of $\overline{Q}_k(x)$ when $D(x) > 0$.

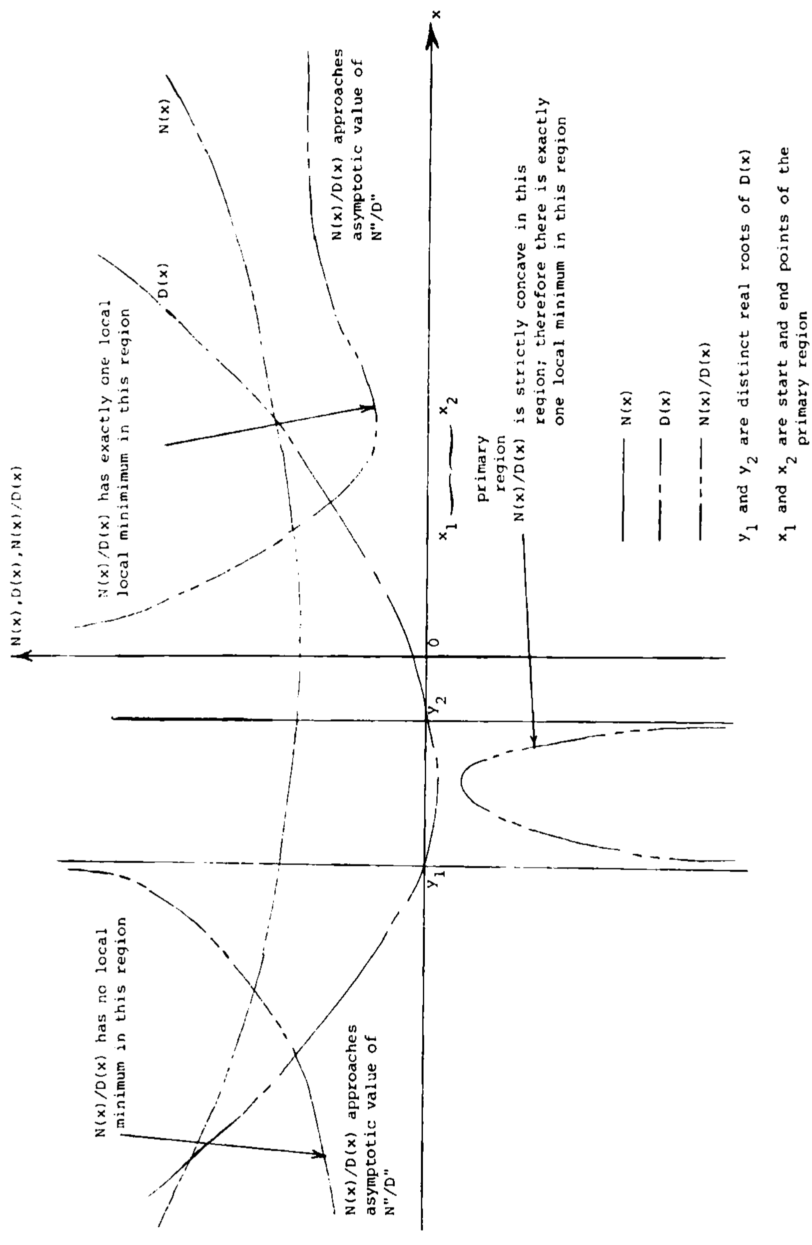

FIG. 4. Illustration for proof of Lemma 1, Subcase (c): Sub-subcase (i).

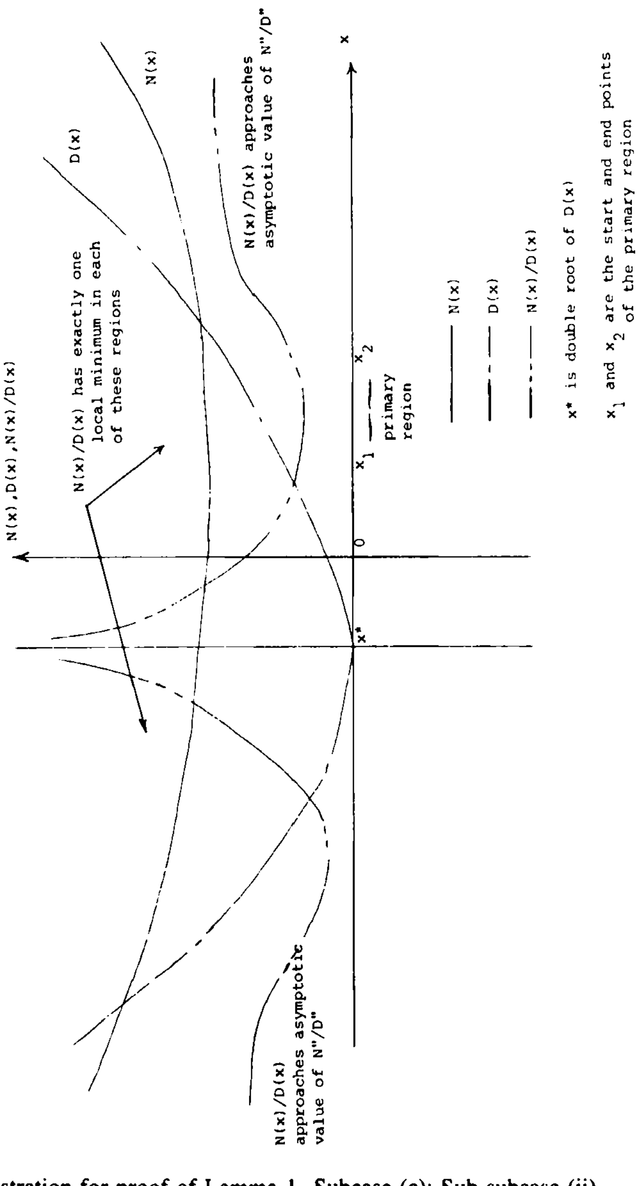

FIG. 5. Illustration for proof of Lemma 1, Subcase (c): Sub-subcase (ii).

Sub-subcase (ii): roots of D(x) are not distinct (Fig. 5). Let the double root of $D(x)$ be x^*. Then exactly one solution to $\overline{Q}_k'(x) = 0$ lies in the interval $(-\infty, x^*]$, and one solution in the interval $[x^*, \infty)$.

Strong quasiconvexity of $\overline{Q}_k(x)$ in the intervals $(-\infty, x^*)$ and (x^*, ∞) follows from the argument in Sub-subcase (i).

The lemma follows. ■

Proof of Theorem 4. (a) The proof is by contradiction. Suppose $\exists\ x_0 \notin T$ and $y_0 \in T$, such that $\overline{TR}(x_0) < \overline{TR}(y_0)$, for $0 \le \lambda \le \varepsilon_1$, for every $\varepsilon_1 > 0$.

Initially pick $\varepsilon_1 = \lambda_{max}/2 > 0$. Then $\overline{TR}(x)$ is finite for $0 \le \lambda \le \varepsilon_1$, and its expression becomes:

$$\overline{TR}(x) = \sum_{k=1}^{K} \pi_k \left[\frac{\frac{\lambda}{2}\overline{S^2}(x)}{\left[1 - \lambda \sum_{j=1}^{k-1} F_j \overline{S}_j(x)\right]\left[1 - \lambda \sum_{j=1}^{k} F_j \overline{S}_j(x)\right]} \right] + \bar{t}_k(x). \tag{11}$$

We consider two cases.

$x_0 \in S - T$. Here $x_0, y_0 \in S$, and $\Sigma_{k=1}^{K} \pi_k \bar{t}_k(x_0) = \Sigma_{k=1}^{K} \pi_k \bar{t}_k(y_0)$.

Upon substituting $a = \overline{S^2}(x_0)/2$, $b_k = \Sigma_{j=1}^{k-1} F_j \overline{S}_j(x_0)$,

$c_k = \Sigma_{j=1}^{k} F_j \overline{S}_j(x_0)$,

$d = \overline{S^2}(y_0)/2$, $e_k = \Sigma_{j=1}^{k-1} F_j \overline{S}_j(y_0)$ and

$g_k = \Sigma_{j=1}^{k} F_j \overline{S}_j(y_0)$, we get:

(note that $a, b_k, c_k, d, e_k, g_k > 0\ \forall\ k = 1, 2, \ldots, K$) $\overline{TR}(x_0) - \overline{TR}(y_0)$

$$= \lambda \Sigma_{k=1}^{K} \pi_k \left[\frac{(a-d) + \lambda[-a(e_k + f_k) + d(b_k + c_k)] + \lambda^2[ae_kf_k - db_kc_k]}{[1-\lambda b_k][1-\lambda c_k][1-\lambda e_k][1-\lambda g_k]} \right] \tag{12}$$
$$= \lambda l(\lambda) < 0.$$

We note that $a > d$, $1 - \lambda b_k > 0$, $1 - \lambda c_k > 0$, $1 - \lambda e_k > 0$, $1 - \lambda g_k > 0$ and $\pi_k > 0$, for $k = 1, 2, \ldots, K$. Therefore (12) leads to a contradiction when ε_1 is chosen small enough, $0 < \varepsilon_1 \le \lambda_{max}/2$ and $0 \le \lambda \le \varepsilon_1$.

$x_0 \notin S - T$. In this case, $x_0 \notin S$, and $\Sigma_{k=1}^{K} \pi_k \bar{t}_k(x_0) > \Sigma_{k=1}^{K} \pi_k \bar{t}_k(y_0)$.

Using notation introduced earlier, we can write:

$$\overline{TR}(x_0) - \overline{TR}(y_0) = \sum_{k=1}^{K} \pi_k \bar{t}_k(x_0) - \sum_{k=1}^{K} \pi_k \bar{t}_k(y_0) + \lambda l(\lambda) < 0. \tag{13}$$

Let $p^* = \min_{0 \le \lambda \le \varepsilon_1} l(\lambda)$, and $s^* = \min(-1, p^*)$.

$$\text{If } \varepsilon_1 = \max\left[\frac{-\left[\sum_{k=1}^{K} \pi_k \bar{t}_k(x_0) - \sum_{k=1}^{K} \pi_k \bar{t}_k(y_0)\right]}{2s^*}, \frac{\lambda_{\max}}{2}\right] > 0,$$

Then (12) is a contradiction for $0 \leq \lambda \leq \varepsilon_1$.

(b) The proof is once again by contradiction. Suppose $\exists\, x_0 \not\in T_1$ and $y_0 \in T_1$, such that $\overline{TR}(x_0) < \overline{TR}(y_0)$, for $\lambda_{\max} - \varepsilon_2 \leq \lambda < \lambda_{\max}$, for every $\varepsilon_2 > 0$.

Initially pick $\varepsilon_2 = \lambda_{\max}/2 > 0$. Then $\overline{TR}(x)$ is as in (11). We consider two cases.

$x_0 \in S_1 - T_1$.

Let

$$g(\lambda) = \sum_{k=1}^{K-1} \pi_k \left[\frac{(a-d) + \lambda[-a(e_k + f_k) + d(b_k + c_k)] + \lambda^2[ae_kf_k - db_kc_k]}{[1-\lambda b_k][1-\lambda c_k][1-\lambda e_k][1-\lambda g_k]}\right],$$

$$L = \max_{0 \leq \lambda < \lambda_{\max}} \lambda\, g(\lambda) + \sum_{k=1}^{K} \pi_k \bar{t}_k(x_0) - \sum_{k=1}^{K} \pi_k \bar{t}_k(y_0), \text{ and}$$

$$L^* = \max(1, L).$$

Then $\overline{TR}(x_0) - \overline{TR}(y_0) < 0$ implies that

$$L^* + \pi_K \left[\frac{\lambda a}{[1-\lambda b_K][1-\lambda c_K]} - \frac{\lambda d}{[1-\lambda e_K][1-\lambda g_K]}\right] < 0. \tag{14}$$

Since $x_0 \in S_1 - T_1$, we have $c_K = g_K$, therefore (14) reduces to:

$$L^* + \frac{\lambda \pi_K}{1-\lambda c_K}\left[\frac{a}{[1-\lambda b_K]} - \frac{d}{[1-\lambda e_K]}\right] < 0. \tag{15}$$

Since $x_0 \in S_1 - T_1$, $a/(1 - \lambda_{\max} b_K) > d/(1 - \lambda_{\max} e_K)$, which implies $a/(1 - \lambda b_K) > d/(1 - \lambda e_K)$, $0 \leq \lambda < \lambda_{\max}$. Let $T^* = a/(1 - \lambda_{\max} b_K/2) - d/(1 - \lambda_{\max} e_K/2) > 0$. Then (15) reduces to:

$$L^* + \frac{\lambda \pi_K}{1-\lambda c_K} T^* < 0. \tag{16}$$

Set $\varepsilon_2 = \lambda_{\max} - \max[\lambda_{\max}/2, L^*/(L^* c_K - \pi_K T^*)] < \lambda_{\max}$, where the inequality follows because $\pi_K, T^* > 0$. We note that $\varepsilon_2 > 0$. For $\lambda_{\max} - \varepsilon_2 \leq \lambda < \lambda_{\max}$, (16) yields a contradiction.

$x_0 \not\in S_1 - T_1$. Pick $\delta_2 = 1/c_K < 1/g_K = \lambda_{\max}$. Then $\overline{TR}(x_0)$ becomes $+\infty$ for $\delta_2 \leq \lambda < \lambda_{\max}$, whereas $\overline{TR}(y_0)$ is finite. Therefore $\overline{TR}(x_0) - \overline{TR}(y_0) < 0$ would yield a contradiction for $\lambda_{\max} - \varepsilon_2 \leq \lambda < \lambda_{\max}$, where $\varepsilon_2 = \lambda_{\max} - \delta_2$.

(c) The proof of this follows from the following facts:

(i) $\overline{S^2}(x)$ is strictly convex on any path of a tree, and on a primary region of a cyclic network.

(ii) $\overline{S^2}(x)/(1 - \lambda_{\max} \Sigma_{k=1}^{K-1} F_k \overline{S}_k(x))$ is strictly convex (when finite) on any path of a tree, and on a primary region of a cyclic network.

(iii) For a tree network if $x,y \in S$, then every point on the path $p[x,y]$ joining x and y belongs to S. The same applies for S_1.

(iv) S and S_1 consist of a finite collection of connected intervals on a cyclic network.

The theorem follows. ∎

References

[1] R. Batta, Facility location in the presence of congestion. Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts (1984).

[2] O. Berman and R. C. Larson, Facility network districting in the presence of queueing. *Transportation Sci.* **19** (1985) 261–277.

[3] O. Berman, R. C. Larson, and S. Chiu, Optimal server location on a network operating as an M/G/1 queue. *Operations Res.* **33** (1985) 746–771.

[4] S. Chiu, Location problems in the presence of queueing. Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts (1982).

[5] S. Chiu, O. Berman, and R. C. Larson, Locating a mobile server queueing facility on a tree network. *Management Sci.* **31** (1985) 764–772.

[6] A. J. Goldman, Optimal locations for centers in a network. *Transportation Sci.* **3** (1969) 352–360.

[7] S. L. Hakimi, Optimal location of switching centers and the absolute centers and medians of a graph. *Operations Res.* **12** (1964) 450–459.

[8] L. Kleinrock, *Queueing Systems, Volume I: Theory*. Wiley, New York (1976).

[9] R. C. Larson, A hypercube queueing model for facility location and redistricting in urban emergency services. *Comput Operations Res.* **1** (1976) 67–95.

[10] J. F. Shapiro, *Mathematical Programming: Structures and Algorithms*. Wiley, New York (1979).

Received May, 1985.
Accepted June, 1987.